National Institute of Environmental Health Sciences

Environmental Factor, July 2013

Bringing down the Tower of Babel in data sharing

By Eddy Ball

As presentations by scientists attending a pair of workshops in June made clear, talking about data sharing is one thing - getting it right is something else entirely.

A workshop co-sponsored by the Office of Scientific Information Management (OSIM) at NIEHS and the Office of Science Information Management (OSIM) at the U.S. Environmental Protection Agency (EPA) June 25 addressed common language concerns in environmental research, while a two-day meeting June 26-27 at NIEHS addressed operability issues that make the task of integrating databases so challenging (see related story).

No one mentioned the biblical story of the Tower of Babel during the workshop on Advancing Environmental Health Data Sharing and Analysis: Finding a Common Language, held at the EPA conference center in Research Triangle Park (RTP), N.C. But the difficulty of developing a consistent nomenclature, among the many now in use, to guide searches of multiple databases was the theme of each of the day's ten presentations.

As the directors of NIEHS OSIM, Allen Dearry, Ph.D., and EPA OSIM, Jerry Blancato, Ph.D., explained in opening remarks, the foundation of effective data sharing is achieving standard language for computerized searches of massive data repositories, to make research data funded by the government publicly available. Data sharing, they emphasized, is an outcome that is not only desirable for research and regulatory scientists, but also one mandated by executive order.

"We've come to a new game in town," said Blancato. "We have to have some type of common language."

A common language emerging from a common ontology

Although the workshop substituted several terms that express the idea in plainer language, some presenters lapsed into database shoptalk with a more comprehensive philosophical term, ontology. Ontology refers to the formulation of definitions, classifications, and relationships, using the tools of logic and formal semantics, in order to most effectively achieve the goal of connecting data across different databases, and make these data accessible to standard software tools.

Unfortunately, as each of the presenters noted, many databases have emerged independently through good faith efforts to meet discipline-specific needs, using terms that may mean one thing for searches of that database, but something different in other contexts.

EPA Information Management Manager Lynne Petterson, Ph.D., offered a telling example of how this ambiguity might affect environmental health research. The word "flow," she explained, means something different to atmospheric physicists than it does to hydrologists - a clash of ontologies that reduces the usefulness of information from their respective databases.

The search for solutions

Like several of her co-presenters at the workshop, Petterson is actively involved in developing what she called "a vocabulary for all seasons," to represent these multiple perspectives, and reconcile past and present meanings of search terms.

Another effort underway at RTI, headed by Carol Hamilton, Ph.D., is developing consensus measures for exposures and biological outcomes, or phenotypes, for use in the NIH Common Data Element Resource Portal, to facilitate genome-wide association studies.

Energizing the push for a common language

While the meeting included a few representatives of databases and data sharing projects elsewhere, the bulk of attendees were people working in the RTP area, representing NIEHS, EPA, the University of North Carolina at Chapel Hill (UNC), RTI International, and North Carolina State University.

This serendipity - the timely coming together of people with a shared interest working within less than an hour's drive of one another - suggested at least one part of a solution for moving common language closer. As part of their efforts to publicize the need for cooperative efforts to build a common language for data sharing, NIEHS Data Scientist Rebecca Boyles noted, concerned scientists can start at home, by getting to know their neighbors, and building a model capable of being replicated and expanded nationwide and even worldwide.

Other suggestions included creating a listserv to extend the conversation, identifying other stakeholders in RTP and elsewhere, investigating other databases with the specific objective of discovering what ontologies they use, and meeting regularly as a group. Several attendees expressed an interest in raising visibility through publications in appropriate journals, such as Environmental Health Perspectives, and trade publications.

Ontology also has important implications for regulatory science. EPA Acting Chief of the Hazardous Pollutant Assessment Group Lyle Burgoon, Ph.D., described his team's work developing a semiautomated predictive tool, for inferring the potential hazards to human health from the thousands of chemicals with insufficient toxicologic value data. This is critical, he said, "Because we can't regulate chemicals with no tox values."

"I didn't realize we had such good resources just down the road," observed workshop facilitator Boyles, data scientist with NIEHS OSIM. (Photo courtesy of Steve McCaw)

Next steps

During the concluding session of the workshop, participants split into small groups for discussion. The groups were charged with brainstorming responses to questions about moving the conversation forward among people with interests in broader and more effective data sharing.

Everyone seemed to agree with what North Carolina State University professor and developer of the NIEHS-funded Comparative Toxicogenomics Database, Carolyn Mattingly, Ph.D., said during her presentation, "You need better ways of navigating the data." The question she and her colleagues faced at the close of the day, however, was just how to gather the momentum for unified progress on a much wider scale (see text box).

"What we are doing is part of that larger effort," Blancato said of the government-wide initiative to improve public access to data. (Photo courtesy of Steve McCaw)

Petterson said her group is striving for an ontology of related-to-ness rather than what-ness. "It [this vocabulary] allows the application to do the work," she said. (Photo courtesy of Steve McCaw)

Despite the challenges his team is facing in bringing order to billions of chemical and exposure data points, like Petterson, Burgoon remains optimistic about the outcome. "I think this is possible," he said. "I think this will actually work." (Photo courtesy of Steve McCaw)

If their smiles are any indication, Dearry, center, and Boyles thought the workshop was meeting its twin objectives of informing people about available resources and motivating the group to redouble its efforts in the quest for a standard ontology. (Photo courtesy of Steve McCaw)

Mattingly identified one major hurdle ahead - convincing database administrators that a common language will benefit them and compensate for the work ahead. "People aren't interested in data entry," she said. (Photo courtesy of Steve McCaw)

During the wrap-up session, Jennifer Fostel, Ph.D., NTP scientific administrator of the Chemical Effects in Biological Systems (CEBS) database, set the tone for future collaborations. "We will publish any data you want to publish," she said. (Photo courtesy of Steve McCaw)

The Environmental Factor is produced monthly by the National Institute of Environmental Health Sciences (NIEHS)
(http://www.niehs.nih.gov/)
, Office of Communications and Public Liaison. The content is not copyrighted, and it can be reprinted without permission. If you use parts of Environmental Factor in your publication, we ask that you provide us with a copy for our records. We welcome your comments and suggestions.
(bruskec@niehs.nih.gov)